EXTRACTION OF GRAPH INFORMATION BASED ON IMAGE CONTENTS AND THE USE OF ONTOLOGY

Sarunya Kanjanawattana[1] and Masaomi Kimura[2]
[1]*Graduate School*
[2]*Information Science and Engineering*
Shibaura Institute of Technology, 3-5-7 Koto-ku Toyosu, Tokyo 135-8548, Japan

ABSTRACT

A graph is an effective form of data representation used to summarize complex information. Explicit information such as the relationship between the X- and Y-axes can be easily extracted from a graph by applying human intelligence. However, implicit knowledge such as information obtained from other related concepts in an ontology also resides in the graph. As this is less accessible, automatic graph information extraction could prove beneficial to users. In this study, we proposed a novel method for extracting both explicit and implicit knowledge from graphs. This was based on our ontology that uses essential information pertaining to the graph and sentence dependency parsing. We focused on two graph types: bar graphs and two-dimensional (2D) charts. Different graph types require different extraction methods and have different extractable features. From the bar graph, we extracted axis labels, the global trend in the data, and the height of the bars. From the 2D charts, we additionally obtained local trends and regression types. The objective was to propose a method for acquiring the implicit and explicit information available in the graphs and entering this into our ontology. For evaluation purposes, we simulated an inquiry involving five questions. Accurate answers were retrieved and significant results were achieved by the shared concepts used in our ontology.

KEYWORDS

Graph information extraction; Ontology; Optical character recognition; Natural language processing

1. INTRODUCTION

Data reported in the academic literature is presented in many formats, including both digital and hard copy. Although readers must read the literature extensively to comprehend the data, its conclusions may be unclear if only descriptive details are available. Graphs are a form of data representation that help readers analyze and extract the information they need, making understanding easier. In a previous study (Kanjanawattana and Kimura, 2015), we attempted to interpret explicit and implicit information in a graph based on a strong relationship between the X- and Y-axes labels and by using information extracted not only from the axis labels themselves but also from data section. The information provided by the axis labels includes implicit knowledge; although not presented directly in the graph, it can be extracted by applying ontology. Human readers find it easier to interpret explicit information presented in a graph; comprehending implicit information is more difficult. A system that allows information to be extracted from a graph can therefore be expected to provide a powerful new approach to knowledge acquisition.

The capture of image semantics has opened up a new field of study that integrates several disciplines to address problems such as semantic gaps (Deserno et al., 2009; Mezaris et al., 2003). To enrich the semantics available from the graph images, we introduced a solution that minimizes the existing problem by using both textual and graphical content from the graph. Several previous studies have focused on the extraction of information based on graph components (Kanjanawattana and Kimura, 2016; Kataria et al., 2008) and the context of the graph (Huang et al., 2005). Kataria et al. (2008) introduced a method for automatically extracting graph elements including data points, axis labels, and legends, and addressed the problem of the overlap between text and data points. Each graph component plays a different role in data interpretation. For example, the axis labels represent a strong relationship and the data points provide real data correlated with the axes. Previous studies used information from these components. However, the focus has been on the

extraction of data components rather than the semantics of the image. Huang et al. (2005) attempted to associate the recognition of textual and graphical information within two-dimensional graphs and captured the semantic content conveyed by graph images. Their approach was to individually identify the texts and graphics in the input image and then combine the extracted information to develop a complete understanding.

The image data used in this study was a collection of line, plot, and bar graphs from journal articles. A bar graph represents the data as bars with lengths proportional to their values. Line and plot graphs are called two-dimensional (2D) charts in this study. We selected graphs containing only single data sets to simplify interpretation.

In this study, we proposed a novel method for extracting the explicit and implicit information present in the data part of the graph. We used a combination of techniques, including ontology, optical character recognition (OCR), and natural language processing (NLP). We addressed the core problem of the semantic gap by making use of both the context of the graph based on the wider document and the graphical content of the graph itself. The objectives of the study were automatic extraction of hidden information using ontology, including the interpretation of explicit information extracted from the data within the graph, and creating ontology of graph information. Our intelligent system offers social benefits, as it can give access to implicit knowledge. It has a range of applications, for example in image interpretation and image search systems. A novelty of the study is that our method was able to extract useful information from the data section of the graphs as well as obtain explicit and implicit information from the relationships within the graph.

2. METHODOLOGY

2.1 Ontology

The ontology used was an extension of that in a previous study (Kanjanawattana and Kimura, 2016). As shown in Figure 1, it supports not only sentence dependency parsing but also graph components and data extracted from graphs. Protégé was used to build the RDF files expressing the ontology. We had already tested its reasoner to validate the generated ontology.

Our ontology included 26 classes and many relations. The main class was the GRAPH class, representing the concept of images from the graph. We used the TYPE class to identify the type of the graph such as bar graph or 2D chart. The 2D chart represents two different graph types: line and plot. We merged these into a single type because of their similar characteristics. Lines in a line graph are formed by combining a large number of plotted points.

Most images were described by their captions and optionally by links to paragraphs. These were represented as CAPTION and PARAGRAPH classes, respectively, and were related to a TOKEN class that stored the concepts of the tokens. Our system assigned part-of-speech (POS) tags and named entity recognition (NER) to each token. We also created dependency relations to represent a typed dependency connecting the tokens in a sentence such as determiner (det) and nominal subject (nsubj).

We identified the basic graph components of axis labels and legends because all graphs use these to represent significant information. For example, the legends of the X- and Y-axes show the relationship between two dimensions. These were therefore made a central part of our ontology. The GRAPH class was related to the COMPONENTS class by a HAS property. The COMPONENTS class comprised three subclasses: X-TITLE, Y-TITLE, and LEGEND. Note that we used only graphs presenting a single data set so that the legend, which shows data labels, was not always essential.

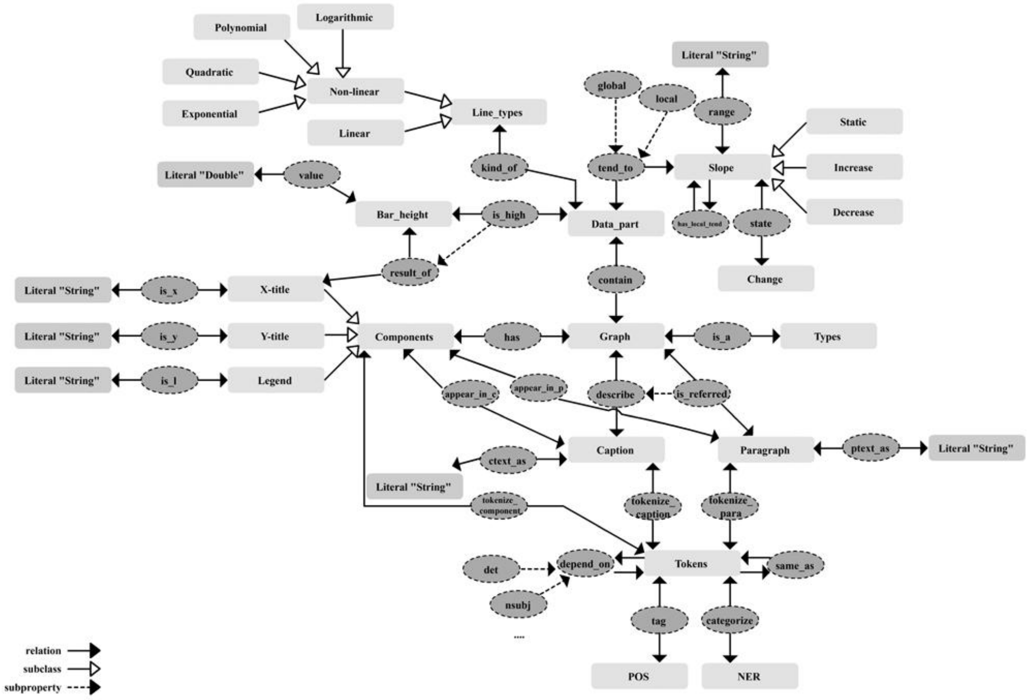


Figure 1. Representation of our ontology structure describing classes, properties, and relations

The real information appears in the data presented in the graph and was recorded as a DATA_PART class. This part of the graph displays a graphical representation of the data, for example by the height of the bar or the slope of the line. The data in a bar graph is represented by rectangular bars corresponding to the categories shown in the X-axis title. A BAR_HEIGHT class was introduced to represent the bar height. 2D charts use plots to show statistical data in a dimensional space. Our approach explored the types of lines used (e.g., linear or non-linear) to represent the data in the graph. This helped predict unseen directions in the data and provide new information that was not described in the caption and paragraphs. We also analyzed and collected both global and local tendencies in a SLOPE class comprising three different trends: an increase (INCREASE class), a decrease (DECREASE class), and no change (STATIC class). The global tendency represents the overall trend in the data while the local tendency provides information about where and how the trend changes. These concepts were described in a CHANGE class.

2.2 Extraction of Graph Information

The core of our study was the introduction of an effective method for extracting significant information from the data part of the graph, including the graph components, and adding this to our ontology. Our proposed method had two steps.

2.2.1 Data Content Identification

We first identified the existing graph components (e.g., X-axis title, Y-axis title, and optionally the legend), including the actual data. As different types of graph provide different information, our system needed a method for analyzing information from each type.

The features generally used for interpreting a bar graph are the X-axis title, the Y-axis title, the height of the bars, and a global tendency corresponding to the centers of the bars. To extract the graph components, the graph image must be partitioned horizontally to acquire the X-axis title and vertically to acquire the Y-axis title. We used OCR to recognize these. However, the occasional presence of irrelevant information such as parts of the bars or numbers may cause misrecognition by the OCR. To address this, we applied a method of automatic graph component extraction described in our previous study (Kanjanawattana and Kimura, 2016).

This method uses a technique of pixel projection to obtain a horizontal profile and remove unnecessary information. This provided cleaned graph components. To interpret bar graphs, we analyzed the height of the bars and the categories on the X-axis.

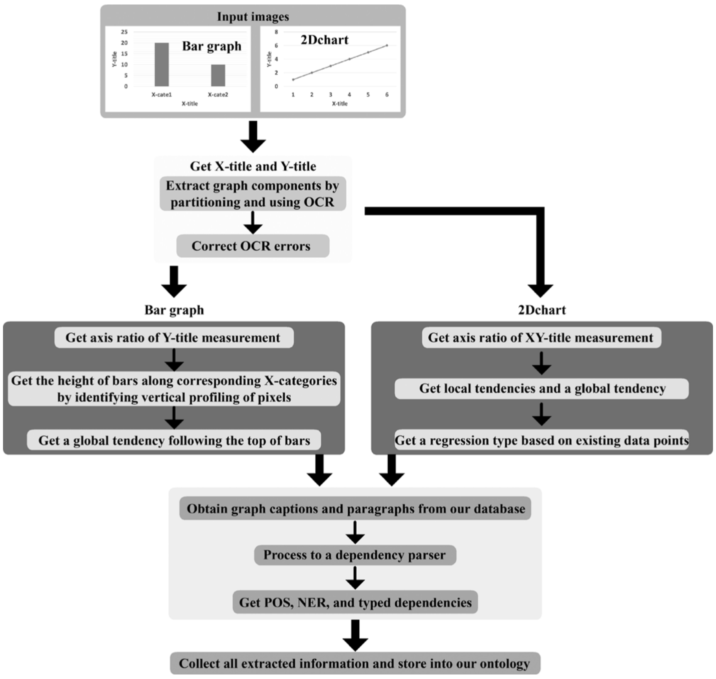


Figure 2. Bar height extraction based on pixel projection and a step function

Our system was able to extract the height of the bars automatically, as shown in Figure 3. After acquiring the cleaned X-axis legend, we used pixel projection with vertical profiling to locate the positions of the bars and their labels. Note that the position of the bars and the labels correspond. When identifying the height of the bars, we applied a step function to smooth the results of the pixel projection and find the center of each bar. We then measured a specific range, equal to half the distance between two neighboring centers, which independently covered each center; we then identified the value of the highest peak within the range. Finally, the graphical bar heights were acquired. However, these values do not match the true scale of the bars, because the proportion of pixels used in each graph varies depending on the data presented. Therefore, the actual bar height must be computed by multiplying the pixel proportion.

We introduced the two-step method of calculating the pixel proportion shown in Figure 4; the steps are data preparation and Y-scale measurement. For data preparation, we initially selected the leftmost partition containing both the Y-axis title and axis measurement after partitioning the graph image. The Y-axis title is irrelevant to the pixel proportion and only the measurement part was retained. Numbers and their respective positions were recognized using OCR. The next step was Y-scale measurement. We obtained the position of each result identified by OCR and measured the difference between two neighboring recognitions, including the difference in vertical distance. We then divided the difference between the two neighbor recognitions by the difference in vertical distance to obtain the actual number of scale units per pixel. We were then able to calculate the actual value of bars by multiplying the height of the bars with the scale units obtained. The global tendency was analyzed from the centers of the bars by calculating the slope.

The main feature of a 2D chart is a line or group of data points. We therefore analyzed the graph components, the global and the local tendencies as well as the regression type. The extraction process for a 2D chart was the same as that for a bar graph component. We initially neglected the titles of both axes to capture the data part. We converted the image to pixel values representing data points in the graph. The global tendency was identified using a global slope derived from the data points. We also attempted to perform a regression analysis using a mathematical library and identify the type of regression that was best suited to the data points using the smallest squared error. Both linear and non-linear regressions were used, including logarithmic, polynomial, quadratic, and exponential regressions.

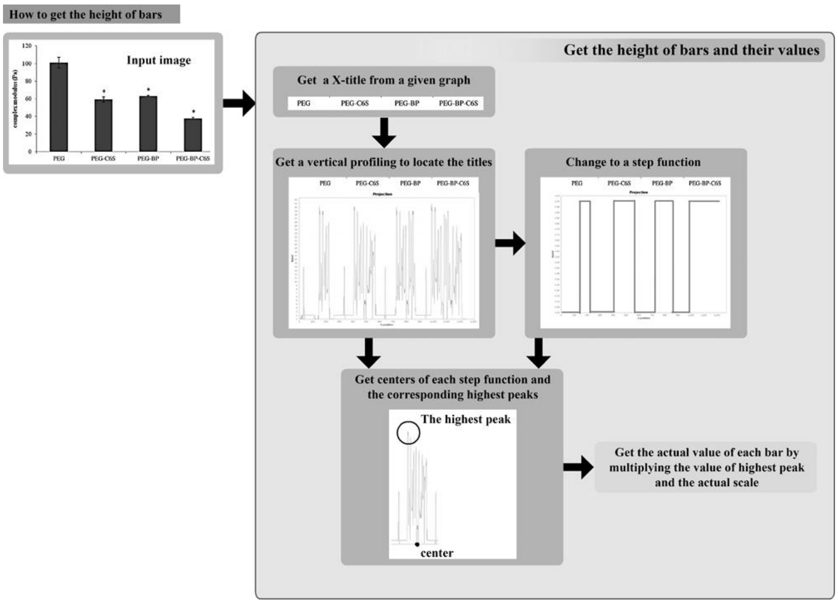


Figure 3. Bar height extraction using pixel projection and a step function

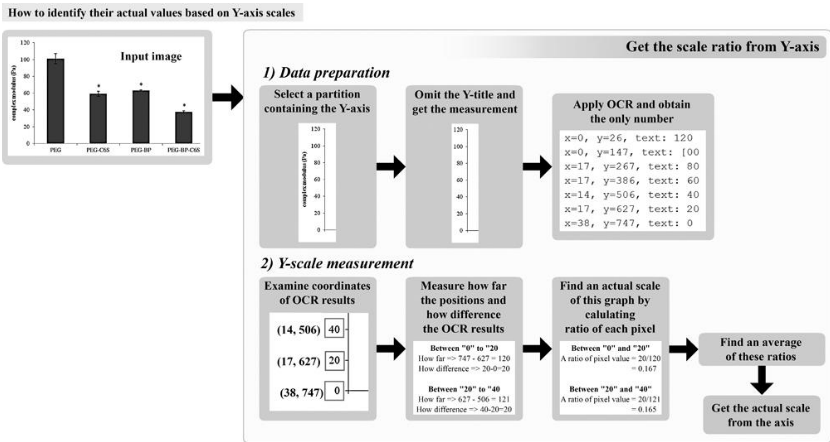


Figure 4. Pixel proportion calculation

A discontinuity in the slope may represent critical information. For example, a line graph may show the oxidation of a chemical substance against temperature and time, while a slope change indicates the saturation point. In recognition of the importance of such local tendencies, we analyzed the trend at each pair of pixel values. If a change was noted between any pair, the change in slope and the position were recorded.

2.2.2 Ontology Construction

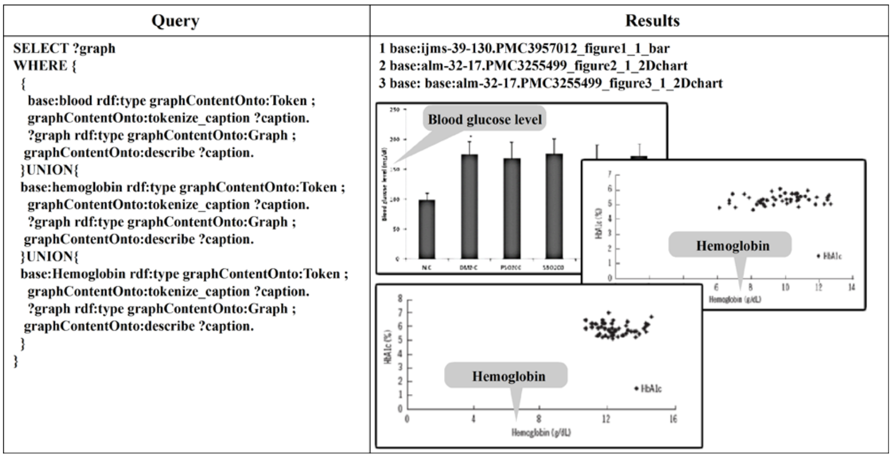
We constructed the classes and relations following our earlier ontology design. The graph contents, such as captions and paragraphs, were stored in a database. These graph descriptions were given in sentences produced by tokenization, as a first step in building the ontology. A dependency parser identified the sentence structures, NER, and POS tags. We endeavored to allocate each word to a category using queries in DBpedia. The queried categories were represented as the NERs of tokens.

ISBN: 978-989-8533-58-6 © 2016

3. EVALUATION AND RESULTS

In this study, the expected outputs were an ontology. Our method provided precise information for the construction of the ontology. To validate our method and ontology, the following questions were applied to the ontology:

1. What graph are both “blood” and “Hemoglobin” related to?
2. How do aphid populations impact sugar?
3. What is the tendency of the number of genes related to green fluorescent protein (EGFP) expression?
4. What value of Lipopolysaccharide (LPS) is described in all graphs and what is its relation?
5. What is a relation between Hemoglobin and Hemoglobin A1c (HbA1c)?



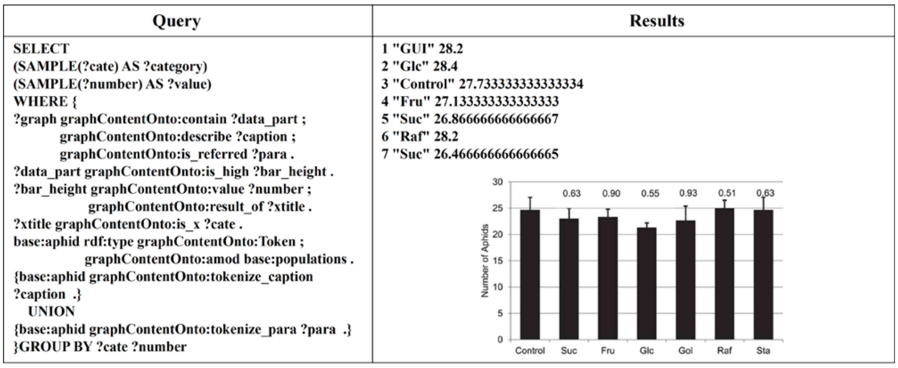
Query

```
SELECT ?graph
WHERE {
 {
  base:blood rdf:type graphContentOnto:Token ;
  graphContentOnto:tokenize_caption ?caption.
  ?graph rdf:type graphContentOnto:Graph ;
 graphContentOnto:describe ?caption.
 }UNION{
 base:hemoglobin rdf:type graphContentOnto:Token ;
  graphContentOnto:tokenize_caption ?caption.
  ?graph rdf:type graphContentOnto:Graph ;
 graphContentOnto:describe ?caption.
 }UNION{
 base:Hemoglobin rdf:type graphContentOnto:Token ;
  graphContentOnto:tokenize_caption ?caption.
  ?graph rdf:type graphContentOnto:Graph ;
 graphContentOnto:describe ?caption.
 }
}
```

Results

1 base:ijms-39-130.PMC3957012_figure1_1_bar
2 base:alm-32-17.PMC3255499_figure2_1_2Dchart
3 base: base:alm-32-17.PMC3255499_figure3_1_2Dchart

Figure 5. SPARQL query command and answers for the first question



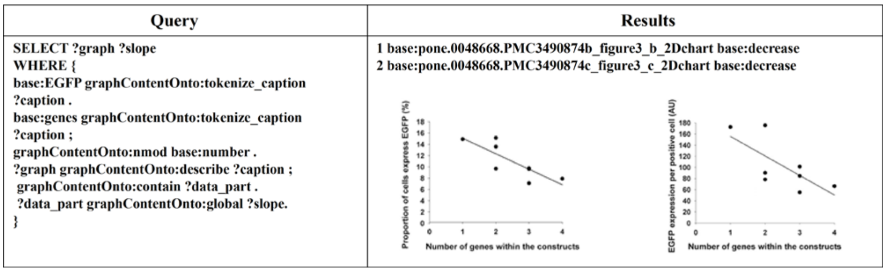
Query

```
SELECT
(SAMPLE(?cate) AS ?category)
(SAMPLE(?number) AS ?value)
WHERE {
?graph graphContentOnto:contain ?data_part ;
        graphContentOnto:describe ?caption ;
        graphContentOnto:is_referred ?para .
?data_part graphContentOnto:is_high ?bar_height .
?bar_height graphContentOnto:value ?number ;
             graphContentOnto:result_of ?xtitle .
?xtitle graphContentOnto:is_x ?cate .
base:aphid rdf:type graphContentOnto:Token ;
            graphContentOnto:amod base:populations .
{base:aphid graphContentOnto:tokenize_caption
?caption .}
  UNION
{base:aphid graphContentOnto:tokenize_para ?para .}
}GROUP BY ?cate ?number
```

Results

1 "GUI" 28.2
2 "Glc" 28.4
3 "Control" 27.733333333333334
4 "Fru" 27.133333333333333
5 "Suc" 26.866666666666667
6 "Raf" 28.2
7 "Suc" 26.466666666666665

Figure 6. SPARQL query command and results for the second question



Query

```
SELECT ?graph ?slope
WHERE {
base:EGFP graphContentOnto:tokenize_caption
?caption .
base:genes graphContentOnto:tokenize_caption
?caption ;
graphContentOnto:nmod base:number .
?graph graphContentOnto:describe ?caption ;
 graphContentOnto:contain ?data_part .
 ?data_part graphContentOnto:global ?slope.
}
```

Results

1 base:pone.0048668.PMC3490874b_figure3_b_2Dchart base:decrease
2 base:pone.0048668.PMC3490874c_figure3_c_2Dchart base:decrease

Figure 7. SPARQL query command and answers for the third question

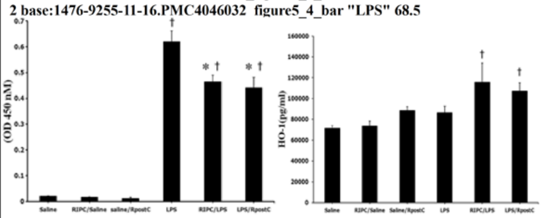
<table>
<tr><th>Query</th><th>Results</th></tr>
<tr><td>SELECT ?graph ?xtitle ?number
WHERE {
base:LPS graphContentOnto:tokenize_component
?component .
?graph graphContentOnto:has ?component.
?component graphContentOnto:is_x ?xtitle.
?bar_heightgraphContentOnto:result_of ?component.
 ?bar_height graphContentOnto:value ?number .
FILTER regex(?xtitle, "^LPS$", "i")
}</td><td>1 base:1476-9255-11-16.PMC4046032_figure4_5_bar "LPS" 79.5
2 base:1476-9255-11-16.PMC4046032 figure5_4_bar "LPS" 68.5
</td></tr>
</table>

Figure 8. SPARQL query command and answers for the fourth question

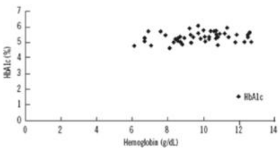
<table>
<tr><th>Query</th><th>Results</th></tr>
<tr><td>SELECT
(SAMPLE(?token) AS ?Subject)
(SAMPLE(?token2) AS ?Predicate)
(SAMPLE(?token3) AS ?Object)
WHERE {
base:hemoglobin graphContentOnto:tokenize_caption
?caption .
base:HbA1c graphContentOnto:tokenize_caption ?caption .
base:A1c graphContentOnto:tokenize_caption ?caption .
?token graphContentOnto:tokenize_caption ?caption.
{?token ?rel ?token2 ;
graphContentOnto:tag base:noun;
?a base:hemoglobin.
}UNION{
?token ?rel ?token2 ;
graphContentOnto:tag base:noun;
?a base:HbA1c.
}UNION{
?token ?rel ?token2 ;
graphContentOnto:tag base:noun;
?a base:A1c.
}
?token2 ?rel2 ?token3 .
?token3 graphContentOnto:tag base:noun.
?rel rdfs:subPropertyOf graphContentOnto:depend_on.
?rel2 rdfs:subPropertyOf graphContentOnto:depend_on.
}GROUP BY ?token ?token2 ?token3</td><td>1 base:hemoglobin base:levels base:plasma
2 base:HbA1c base:levels base:correlation
3 base:hemoglobin base:levels base:correlation
4 base:HbA1c base:levels base:plasma
5 base:HbA1c base:hemoglobin base:A1c
6 base:HbA1c base:hemoglobin base:levels
7 base:HbA1c base:levels base:Correlation
8 base:hemoglobin base:levels base:Correlation
9 base:HbA1c base:A1c base:Abbreviation
10 base:hemoglobin base:A1c base:Abbreviation
11 base:hemoglobin base:levels base:Quantification
12 base:HbA1c base:levels base:Quantification

The caption is "Correlation between hemoglobin and HbA1c levels in patients after 1 month of treatment. Abbreviation: HbA1c, hemoglobin A1c."</td></tr>
</table>

Figure 9. SPARQL query command and answers for the fifth question

Note that all our input graphs were in the field of biology, as the data were collected from journals available through PubMed. SPARQL queries were built to retrieve the related classes and relations of the ontology. The simulation was meant to model a user attempting to use our ontology and deciding what kind of question should be asked.

The SPARQL queries and their results are presented in Figures 5–9. Figure 5 shows the query command and the results obtained for the first question. Three graphs presented by Nekooeian et al. (2014) and Sinha et al. (2012) mentioned “blood” or “Hemoglobin” in their captions. Figure 6 shows a graph by Cao et al. (2013), with the values of each bar representing the impact of aphid populations on sugar. Figure 7 presents answers to the third question from a graph relating the number of genes and the EGFP expression by Gao et al. (2012). Figure 8 shows how our SPARQL query interrogated the ontology to retrieve graphs by Kim et al. (2014) pertaining to LPS and includes its values. Figure 9 shows the results for the correlation between “Hemoglobin” and “HbA1c” in a graph presented by Sinha et al. (2012). This displayed all tokens that had at least one relation with the specified tokens. For quantitative evaluation, we analyzed the precision and observed errors that arose in the course of the simulation. For the aforementioned five questions, we obtained relevant answers by using five queries. However, errors arose due to OCR misrecognition. These were ignored because they were not related to the validity of the ontology.

4. DISCUSSION

In this study, we proposed a new method of extracting information from a graph based on the use of ontology. We extracted the graph components and data located in the data section of the graph. A dependency parser was applied to analyze the captions of the graph and related paragraphs. The category to which each token belonged was acquired from DBpedia. The method was then applied to a graph-based

search engine with user queries in the field of biology. The goal was to use the ontology to extract both implicit and explicit information from the graphs. Five inquiries were run, and the answers returned were, in the main, correct. Unfortunately, in some cases (e.g., the second question), failures in OCR introduced errors. The accuracy of the results provided evidence that our method was able to precisely extract information from the graphs. For the fifth question, answers were found from the captions of the retrieved graphs and several triples representing tokens that were connected by dependencies were obtained. Interestingly, we were able to retrieve tokens that were not available from the captions of the graphs, but were instead taken from other graphs sharing the same concepts such as “quantification” and “plasma.” Based on this result, we believe that our ontology is suitable for use in inquiries involving information pertaining to a graph. However, a limitation of the study was that we focused only on a limited set of graphs: line, plot, and bar. Our system does not yet support analysis of other graph types that require a different method of interpretation. Moreover, the system currently cannot deal with multiple data.

5. CONCLUSIONS

We developed an effective method for extracting graph information, and an ontology to support the dependency parsing of English sentences. Several techniques were combined to achieve this: OCR, NLP, and ontology. We evaluated the method by using the constructed ontology to address five questions. Accurate answers were obtained and significant results were achieved by the shared concepts used in our ontology, thereby demonstrating the effectiveness of the method. In future studies, we will develop the system further by building a simple user interface and extending the dataset to allow quantitative evaluations. We may also extend the domain of search data to other fields such as engineering.

REFERENCES

Cao, Te and Lahiri, Ipsita and Singh, Vijay and Louis, Joe and Shah, Jyoti and Ayre, Brian G, 2013. Metabolic engineering of raffinose-family oligosaccharides in the phloem reveals alterations in carbon partitioning and enhances resistance to green peach aphid. *Frontiers in Plant Science*, 4(2013), p. 263.

Deserno, T.M., Antani, S. and Long, R., 2009. Ontology of Gaps in Content-Based Image Retrieval. *Journal of digital imaging*, Vol. 22, No. 3, pp. 202-215.

Gao, S.Y., Jack, M.M. and O’Neill, C., 2012. Towards optimising the production of and expression from polycistronic vectors in embryonic stem cells. *PLoS ONE*, Vol. 7, No. 11, p. e48668.

Huang, W., Tan, C.L. and Leow, W.K., 2005. Associating text and graphics for scientific chart understanding. *In Eighth International Conference on Document Analysis and Recognition (ICDAR’05)*. IEEE, pp. 580-584.

Kanjanawattana, S. and Kimura, M., 2015. A proposal for a method of graph ontology by automatically extracting relationships between captions and X- and Y-axis titles. In *IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Kanjanawattana, S. and Kimura, M., 2016. Extraction and identification of bar graph components by automatic Epsilon estimation. *The 9th International Conference on Advanced Computer Theory and Engineering (ICACTE 2016).*

Kanjanawattana, S. and Masaomi, K., 2016. Ontologies-based Optical Character Recognition-error Correction Method for Bar Graphs. *The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016).*

Kataria, S. et al., 2008. Automatic Extraction of Data Points and Text Blocks from 2-Dimensional Plots in Digital Documents. *In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008).* pp. 1169-1174.

Kim, Y.-H. et al., 2014. Effect of remote ischemic post-conditioning on systemic inflammatory response and survival rate in lipopolysaccharide-induced systemic inflammation model. *Journal of inflammation (London, England)*, Vol. 11, No. 1, p. 1.

Mezaris, V., Kompatsiaris, I. and Strintzis, M.G., 2003. An ontology approach to object-based image retrieval. *In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. IEEE, Vol. 2, pp. II-511.

Nekooeian, A.A. et al., 2014. Effects of pomegranate seed oil on insulin release in rats with type 2 diabetes. *Iranian Journal of Medical Sciences*, Vol. 39, No. 2, pp. 130-135.

Sinha, N. et al., 2012. Effect of iron deficiency anemia on hemoglobin A1c levels. *Annals of Laboratory Medicine*, Vol. 32, No. 1, pp. 17-22.